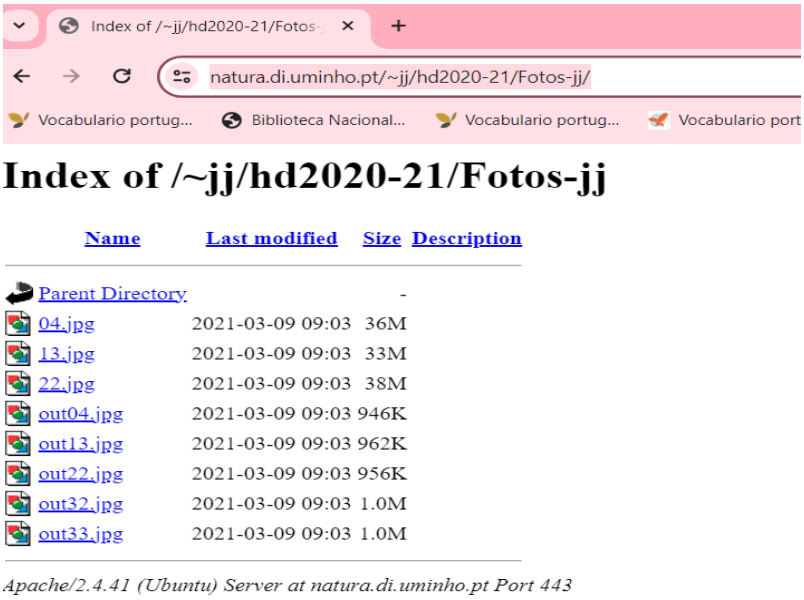
Diário de bordo 16/2

Realizado por: Cristiana Gomes
Piloto de testes: Gabriela Macieira

Antes de tudo, acedemos a este link: https://natura.di.uminho.pt/~jj/hd2020-21/Fotos-jj/ ; que nos encaminhou para um site com algumas fotos do professor. As fotos têm mais de 100 anos e ótima qualidade.



O professor recomendou outro website para visitarmos: humans in NY. https://www.humansofnewyork.com/

Depois disto, demos início à aula. Retomámos o exercício da aula anterior de recorte de artigos (pasta A2). Aprofundámos BeautifulSoup.

O professor mostrou algumas cábulas de bs4:

1. Na imagem seguinte, exemplo de código onde, a partir do beautiful soup, acedemos a determinado url e extraímos as informações em html e xml. item_tags contém a função find_all que, neste caso, vai procurar os elementos que comecem na etiqueta <item> e terminem na </item>.

```
from bs4 import BeautifulSoup as bs

url = "http://...."
dt = bs(requests.get(url).text, 'lxml') ## ('lxml-xml' para xml)
dt = bs(requests.get(url).text, 'html.parser')

item_tags = dt.find_all('item')
                        (text="...")
```

2. Depois, algumas funcionalidades, como:
a.get(“href”): usado para obter o valor do atributo "href"(o link) de <a>.
a.get_attribute_list(“class”): usado para obter uma lista de valores associados ao atributo de classe da tag.
urls = [a.get("href") for a in dt.find_all("a")]: organiza em lista todos os urls encontrados na tag <a>.
.name: dá o nome da tag.

```
for a in dt.find_all("a")
    ... a.get("href")
        .get_attribute_list("class")
urls=[ a.get("href") for a in dt.find_all("a")]

  .contents[0].name
  .children
  .stripped_strings
```

3. Nesta imagem, algumas maneiras de como usar o método find_all.
(Encontrei um site, https://scrapeops.io/python-web-scraping-playbook/python-beautifulsoup-findall/ , onde explica algumas das funcionalidades).

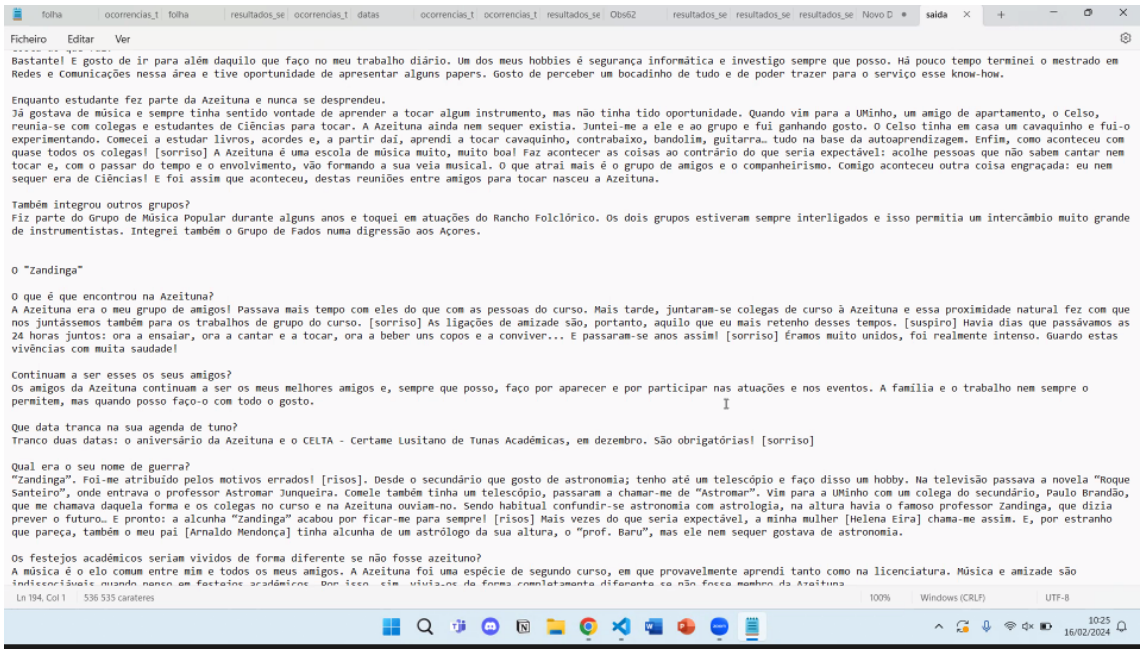
```
find_all
  True          all elements
   "p"
   "a","b"      <a><b>...  "a" with a "b"
   ["a","b"]    tag a ou b
   re.compile(r'\w\wbbb')
   id="v"
   attr={id:val}
   string=re.compile()
   class_ =
   function dt → bool
```

4. Este print tem alguns exemplos de como alterar valores de tags:

```
modificar
   del a[href]
   a.unwrap()      retira a tag, mantem os conteudos
   a.replaceWithChildren
   a[href]=v
   a.name = "f"
   a.string = "v"
   a.appent()  .extend() .insert()
   a.clear()   elimina contents
   a = dt.b.extract()
   a.b.decompose()   elimina a tag b e seu content
natura:hd2020-21$
```

Voltando ao exercício. Ponto da situação: texto sem etiquetas. Guardámos este texto limpo num ficheiro à parte, “saida.txt”

```python
from jjcli import *
from bs4 import BeautifulSoup as bs


ats= glob("nos_1146_1149/Article.aspx*")
print (ats)

fo= open ("saida.txt", "w", encoding="utf-8")

def proc_article(html):
    #print (len(html)) #está a contar os carateres de
    a=bs(html) # cria uma árvore documental
    art= a.find("div", id="artigo") #procura no html
    print ("==========\n", art.get_text(), file= fo) #

for file in ats:
    with open(file, encoding="utf-8") as f:
        html= f.read()
    proc_article (html)

```



Bastante! E gosto de ir para além daquilo que faço no meu trabalho diário. Um dos meus hobbies é segurança informática e investigo sempre que posso. Há pouco tempo terminei o mestrado em Redes e Comunicações nessa área e tive oportunidade de apresentar alguns papers. Gosto de perceber um bocadinho de tudo e de poder trazer para o serviço esse know-how.

Enquanto estudante fez parte da Azeituna e nunca se desprendeu.
Já gostava de música e sempre tinha sentido vontade de aprender a tocar algum instrumento, mas não tinha tido oportunidade. Quando vim para a UMinho, um amigo de apartamento, o Celso, reunia-se com colegas e estudantes de Ciências para tocar. A Azeituna ainda nem sequer existia. Juntei-me a ele e ao grupo e fui ganhando gosto. O Celso tinha em casa um cavaquinho e fui-o experimentando. Comecei a estudar livros, acordes e, a partir daí, aprendi a tocar cavaquinho, contrabaixo, bandolim, guitarra… tudo na base da autoaprendizagem. Enfim, como aconteceu com quase todos os colegas! [sorriso] A Azeituna é uma escola de música muito, muito boa! Faz acontecer as coisas ao contrário do que seria expectável: acolhe pessoas que não sabem cantar nem tocar e, com o passar do tempo e o envolvimento, vão formando a sua veia musical. O que atrai mais é o grupo de amigos e o companheirismo. Comigo aconteceu outra coisa engraçada: eu nem sequer era de Ciências! E foi assim que aconteceu, destas reuniões entre amigos para tocar nasceu a Azeituna.

Também integrou outros grupos?
Fiz parte do Grupo de Música Popular durante alguns anos e toquei em atuações do Rancho Folclórico. Os dois grupos estiveram sempre interligados e isso permitia um intercâmbio muito grande de instrumentistas. Integrei também o Grupo de Fados numa digressão aos Açores.

O "Zandinga"

O que é que encontrou na Azeituna?
A Azeituna era o meu grupo de amigos! Passava mais tempo com eles do que com as pessoas do curso. Mais tarde, juntaram-se colegas de curso à Azeituna e essa proximidade natural fez com que nos juntássemos também para os trabalhos de grupo do curso. [sorriso] As ligações de amizade são, portanto, aquilo que eu mais retenho desses tempos. [suspiro] Havia dias que passávamos as 24 horas juntos: ora a ensaiar, ora a cantar e a tocar, ora a beber uns copos e a conviver... E passaram-se anos assim! [sorriso] Éramos muito unidos, foi realmente intenso. Guardo estas vivências com muita saudade!

Continuam a ser esses os seus amigos?
Os amigos da Azeituna continuam a ser os meus melhores amigos e, sempre que posso, faço por aparecer e por participar nas atuações e nos eventos. A família e o trabalho nem sempre o permitem, mas quando posso faço-o com todo o gosto.

Que data tranca na sua agenda de tuno?
Tranco duas datas: o aniversário da Azeituna e o CELTA - Certame Lusitano de Tunas Académicas, em dezembro. São obrigatórias! [sorriso]

Qual era o seu nome de guerra?
"Zandinga". Foi-me atribuído pelos motivos errados! [risos]. Desde o secundário que gosto de astronomia; tenho até um telescópio e faço disso um hobby. Na televisão passava a novela "Roque Santeiro", onde entrava o professor Astromar Junqueira. Comele também tinha um telescópio, passaram a chamar-me de "Astromar". Vim para a UMinho com um colega do secundário, Paulo Brandão, que me chamava daquela forma e os colegas no curso e na Azeituna ouviam-no. Sendo habitual confundir-se astronomia com astrologia, na altura havia o famoso professor Zandinga, que dizia prever o futuro… E pronto: a alcunha "Zandinga" acabou por ficar-me para sempre! [risos] Mais vezes do que seria expectável, a minha mulher [Helena Eira] chama-me assim. E, por estranho que pareça, também o meu pai [Arnaldo Mendonça] tinha alcunha de um astrólogo da sua altura, o "prof. Baru", mas ele nem sequer gostava de astronomia.

Os festejos académicos seriam vividos de forma diferente se não fosse azeituno?
A música é o elo comum entre mim e todos os meus amigos. A Azeituna foi uma espécie de segundo curso, em que provavelmente aprendi tanto como na licenciatura. Música e amizade são

Decidimos criar uma função extra, que colocámos no proc_article; vimos os metas, vimos os atributos, e formámos um cabeçalho. Exemplo de metas <og:title>. Criámos um loop para ir acrescentando cada um dos metas que precisávamos:

```
9
10  def proc_article(html):
11      #print (len(html)) #está a contar os carateres de cada artigo
12      a=bs(html) # cria uma árvore documental
13      for meta in a.find_all ("meta"):
14          print (meta.get("property"), ":" , meta.get("content"))
15      art= a.find("div", id="artigo") #procura no html
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS Python

```
og:type : blog
None : None
None : IE=EmulateIE9
og:title : "Antes trabalhava-se com mais tempo"
og:url : http://www.nos.uminho.pt/Article.aspx?id=3692
og:image : http://www.nos.uminho.pt/Images/destaques/20231220165645_Conceio14.jpg
og:site_name : Jornal Online UMINHO "Antes trabalhava-se com mais tempo"
og:description : Tem 62 anos e nasceu em Braga, onde vive. A técnica superior começou a trabal
har na UMinho em 1980, ainda a Escola de Letras, Artes e Ciências Humanas era a Unidade Pedagó
gica de Letras e Artes. Já lá vão 43 anos!
```

Tentamos retirar o "og:"; para isso criámos um if- Se o meta fosse vazio (como na imagem de cima, "None: None"), ignorava e continuava para o meta seguinte;

Depois criámos a string "cabeçalho" onde colocámos tudo organizado, que fica da seguinte forma:

```
def proc_article(html): #Define uma função chamada proc_article c
    #print(len(html))  #está a contar os caracteres de cada artig
    a=bs(html)  #cria uma árvore
    cabecalho = ""
    for meta in a.find_all("meta"):
        p = meta.get("property") #Itera sobre todas as tags "meta
        if p is None:
            continue
        p = p.replace("og:", "") #se existir, remove o prefixo "o
        cabecalho+= f"{p}: {meta.get('content')}\n" #constrói uma
        #print(p, ":" , meta.get("content"))
    art = a.find("div", id="artigo")
    print("==========\n", cabecalho, art.get_text(), file = fo)
```

Probelma: perderam-se imagens, que passaram a aparecer como:

<img src=.."
alt=".." />

Tpc:

(1) Procurar extrair e por as datas no cabeçalho
2) Colocar depois do Cabeçalho três (---)
3) Função limpezas (tirar lixo tipo voltar à página anterior)

4) Encontrar a lista das imagens e por no cabeçalho

Extra: Exemplo que o professor colocou no quadro; quando temos duas pastas diferentes, e queremos incluir fotos de uma específica: glob (“P* / *.jpg”) ; se quero incluir fotos de P1 e P2: glob ("P** / *.jpg , _____)